Amendments to the Specification

Please replace the paragraph at page 3, lines 2-6 with the following amended paragraph:

URL stands for Uniform Resource Locator. Generally, URLs have three parts: the first part describes the protocol used to access the content pointed to by the URL, the second contains the directory in which the content is located, and the third contains the file that stores the content:
<protocol> : <domain> <directory> <file>
where "protocol" may be of the type http, "domain" is a domain name of the directory in which a file so named is located.

Please delete the paragraph at page 3, lines 7 through 11 which starts with "For example:".

Please replace the paragraph at page 3, lines 15 through 19 with the following amended paragraph:

For example, the following are legal variations of ~~the previous example~~ URLs:

www.corex.com/bios.html

www.cardscan.com

fn.cnn.com/archives/may99/pr37.html

~~ftp://shiva.lin.com/soft/words.zip~~

Please replace the paragraph at page 5, lines 1 through 9 with the following amended paragraph:

Decades of active research in the Computer Science field of Information Retrieval have ~~yield~~ yielded several algorithms and techniques for efficiently searching and retrieving information from structured databases. However, the world's largest information repository, the Web, contains mostly unstructured information, in the form of Web pages, text documents, or

multimedia files. There are no standards on the content, format, or style of information published in the Web, except perhaps, the requirement that it should be understandable by human readers. Therefore the power of structured database queries that can readily connect, combine and filter information to present exactly what the user wants is not available in the Web.

Please replace the paragraph at page 5, line 26 through page 6, line 9 with the following amended paragraph:

Examples of some well-known search engines today are Yahoo, Excite, Lycos, Northern Light, AltaVista, Google, etc. Examples of inventions that attempt to extract structured data from the Web are disclosed in sections 5, 6, and 7 of the related U.S. Provisional Application No. 60/221,750 filed on July 31, 2000 for a "Computer Database Method and Apparatus". These two separate groups of applications (search engines and data extractors) have different approaches to the problem of Web information retrieval; however, they both share a common need: they need a tool to "feed" them with pages from the Web so that they can either index those pages, or extract data. This tool is usually an automated program (or, "software robot") that visits and traverses lists of Web sites and is commonly referred to as a "Web crawler". Every search engine or Web data extraction tool uses one or more Web crawlers that are often specialized in finding and returning pages with specific features or content. Furthermore, these software robots are "smart" enough to optimize their traversal of Web sites so that they spend the minimum possible time in a Web site but return the maximum number of relevant Web pages.

Please replace the paragraph at page 13, lines 13 through 24 with the following amended paragraph:

As mentioned above, in step 40 of Fig. 2, one important task that the Crawler 11 performs is to identify the content owner name of every Web site that it visits. Knowing the content owner name is an important piece of information for several reasons:

a) it enables better data extraction from the Web site, since it provides a useful meta-understanding of text found in the site. For example, if the Crawler 11 identifies the

site's owner name as "ABC Corporation", then a list of people found in a paragraph headed "Management Team" can be safely assumed to be employees of "ABC Corporation".

b) it facilitates algorithms for resolving duplicate sites (see below).

c) it creates automatically a list of domain URL[[']]s with corresponding owner name, which is of high business value.

Please replace the paragraph at page 13, line 25 through page 14, line 2 with the following amended paragraph:

In order to identify the content owner name of a Web site, the current invention uses a system based on Bayesian Networks described in ~~Invention 1 as disclosed in~~ section 1 of the related U.S. Provisional Application No. 60/221,750 ~~filed on July 31, 2000 for a "Computer Database Method and Apparatus"~~.